Genet. Sel. Evol. 34 (2002) 275–305

DOI: 10.1051/gse:2002009

Review

A review on SNP and other types of molecular markers and their use in animal genetics

Alain VIGNAL[a]*, Denis MILAN[a], Magali SANCRISTOBAL[a], André EGGEN[b]

[a] Laboratoire de génétique cellulaire, Inra, chemin de Borde-Rouge, Auzeville BP 27, 31326 Castanet-Tolosan cedex, France
[b] Laboratoire de génétique biochimique et de cytogénétique, Inra, domaine de Vilvert, 78352 Jouy-en-Josas cedex, France

(Received 11 February 2002; accepted 8 March 2002)

Abstract – During the last ten years, the use of molecular markers, revealing polymorphism at the DNA level, has been playing an increasing part in animal genetics studies. Amongst others, the microsatellite DNA marker has been the most widely used, due to its easy use by simple PCR, followed by a denaturing gel electrophoresis for allele size determination, and to the high degree of information provided by its large number of alleles per locus. Despite this, a new marker type, named SNP, for Single Nucleotide Polymorphism, is now on the scene and has gained high popularity, even though it is only a bi-allelic type of marker. In this review, we will discuss the reasons for this apparent step backwards, and the pertinence of the use of SNPs in animal genetics, in comparison with other marker types.

SNP / microsatellite / molecular marker / genome / polymorphism

1. INTRODUCTION: OLDER TYPES OF MOLECULAR GENETIC MARKERS

Molecular markers, revealing polymorphisms at the DNA level, are now key players in animal genetics. However, due to the existence of various molecular biology techniques to produce them, and to the various biological implications some can have, a large variety exists, from which choices will have to be made according to purposes.

Two main points have to be considered, when using molecular markers for genetic studies. As seen from the molecular biologist's point of view, the genotyping procedure should be as simple and have as low a cost as possible, in

* Correspondence and reprints
E-mail: vignal@toulouse.inra.fr

order to generate the vast amount of genotyping data often necessary. From the statistician's point of view, according to the type of analysis to be performed, a few characteristics are important, such as the dominance relationships, information content, neutrality, map positions or genetic independence of markers. Whatever the system chosen, the data must of course be as reliable as possible.

From the molecular mechanism point of view, the three main variation types at the DNA level, are single nucleotide changes, now named SNPs for single nucleotide polymorphisms; insertions or deletions (Indels) of various lengths ranging from 1 to several hundred base pairs and VNTR, for variations in the number of tandem repeats (Tab. I). The molecular techniques used for genotyping will be adapted to the variation type and to the scale and throughput envisaged (Tab. II).

If we consider molecular genetic DNA markers in terms of the type of information they provide at a single locus, only three main categories can be described, in increasing degrees of interest: the bi-allelic dominant, such as RAPDs (random amplification of polymorphic DNA), AFLPs (amplified fragment length polymorphism); the bi-allelic co-dominant, such as RFLPs (restriction fragment length polymorphism), SSCPs (single stranded conformation polymorphism) and the multi-allelic co-dominant, such as the microsatellites. Bearing this in mind, some variations in the popularity of the markers used at different periods of time in the recent and quickly evolving field of molecular genetics, can be easily understood.

One of the most dramatic examples, is that of the replacement of RFLPs by microsatellites for building genetic maps in human and animal species. Indeed, the first large scale effort to produce a human genetic map, was performed mainly using RFLP markers, the best known genetic markers at the time [20]. However, with the generalisation of PCR and the demonstration of Mendelian inheritance of the multiple alleles due to variations in the number of short nucleotide repeats observed at microsatellite loci [50, 81], a change in strategy was quickly made and all the successive genetic maps in humans [14, 18, 82] were based mainly on this new type of marker. Two main reasons were behind this quick shift. The first was the high number of alleles present at a single microsatellite locus, leading to high heterozygosity values, therefore enabling to dramatically reduce the number of reference families to be used for building the map. The second was the possibility to perform genotypes by simple PCR, followed by allele sizing on polyacrylamide gels. Microsatellite based maps also exist for species of agricultural interest, with the main ones being the cow [38], pig [67], chicken [27], sheep [53], goat [77], and horse [75].

As for the other marker types, although at a first glance they do not seem that interesting to use, due to the fact that they are of the dominant type, the RAPDs and AFLPs have a great advantage in terms of ease of use in the laboratory. Indeed, fingerprint types of patterns are produced by just using

standard oligonucleotides in combination (in addition to restriction enzymes in the case of AFLPs), considerably reducing the effort and consumables, and therefore the price, needed to produce the genotypes for a large scale study. Once the technique has been set to work in the laboratory, data can be produced for different species by using exactly the same reagents and conditions. However, the drawback is that the markers are generally dominant and generated at random. The dominance problem can be partially overcome by the possibility of quickly generating high density maps and the lack of prior mapping information means that once linkage has been established between markers from a linkage group and a phenotype, the work will focus only on that one particular region, leaving the rest of the genome aside. One major problem with the RAPDs, is their low reproducibility, depending highly on the PCR conditions. Contrariwise, AFLP markers can still be a good choice for QTL mapping or diversity studies in species devoid of dense marker maps [78].

After a whole decade of domination in the molecular genetics field for human and animal genome studies by the microsatellite markers, a new type of marker, named SNP (single nucleotide polymorphism), recently appeared on the scene. To have a better prospect on the implications they have, we will describe SNPs together with the methods used for producing and genotyping them. Comparisons with other types of markers will be done, as a guideline to the markers to be chosen according to the various types of studies envisaged.

2. SNPS

2.1. Definition of SNPs and the generation of single nucleotide polymorphisms

As suggested by the acronym, an SNP (single nucleotide polymorphism) marker is just a single base change in a DNA sequence, with a usual alternative of two possible nucleotides at a given position. For such a base position with sequence alternatives in genomic DNA to be considered as an SNP, it is considered that the least frequent allele should have a frequency of 1% or greater. Although in principle, at each position of a sequence stretch, any of the four possible nucleotide bases can be present, SNPs are usually bi-allelic in practice. One of the reasons for this, is the low frequency of single nucleotide substitutions at the origin of SNPs, estimated to being between 1×10^{-9} and 5×10^{-9} per nucleotide and per year at neutral positions in mammals [48,57]. Therefore, the probability of two independent base changes occurring at a single position is very low. Another reason is due to a bias in mutations, leading to the prevalence of two SNP types. Mutation mechanisms result either in transitions: purine-purine (A $\Leftrightarrow$ G) or pyrimidine-pyrimidine (C $\Leftrightarrow$ T) exchanges, or transversions: purine-pyrimidine or pyrimidine-purine

(A ⇔ C, A ⇔ T, G ⇔ C, G ⇔ T) exchanges. With twice as many possible transversions than transitions, the transitions over transversions ratio, should be 0.5 if mutations are random. However, observed data indicate a clear bias towards the transitions. For instance a study comparing rodent and human sequences indicates a transition rate equal to 1.4 times that of transversions, implying that each type of transitional change is produced 2.8 times as often as each type of transversional change [11]. More recent results obtained from a study of human SNPs from EST sequence trace databases gave a transition to transversion ratio of 1.7 [63]. The results obtained to date in chickens indicate higher ratios than in mammals: SNPs mined from EST sequence traces gave a ratio of 2.3 [74] or 4 [39] and a survey of 138 SNPs from non-coding DNA in chickens gave a ratio of 2.36 (Vignal and Weigend, unpublished data). One probable explanation for this bias is the high spontaneous rate of deamination of 5-methyl cytosine (5mC) to thymidine in the CpG dinucleotides, leading to the generation of higher levels of C ⇔ T SNPs, seen as G ⇔ A SNPs on the reverse strand [13, 80].

Some authors consider one base pair indels (insertions or deletions) as SNPs, although they certainly occur by a different mechanism.

2.2. SNPs: a new type of molecular marker?

What is the reason for the increasing popularity of SNPs, whereas in terms of genetic information provided, as simple bi-allelic co-dominant markers, they can be considered as a step backwards when compared to the highly informative multi-allelic microsatellites? Are we not only putting a new name on what has just been considered until now as a common polymorphism and originally studied as RFLPs? In fact, the more recent SNP concept has basically arisen from the recent need for very high densities of genetic markers for the studies of multifactorial diseases, and the recent progress in polymorphism detection and genotyping techniques.

3. SNP DISCOVERY

3.1. Principal strategies

Although numerous approaches for SNP discovery have been described, including some also currently used for genotyping, the main ones are based on the comparison of locus-specific sequences, generated from different chromosomes. The simplest, when targeting a defined region for instance containing candidate genes, is to perform direct sequencing of genomic PCR products obtained in different individuals. However, on a large scale, this approach tends to be costly due to the need for locus-specific primers, is limited to regions for which sequence data is available, and produces a diploid sequence

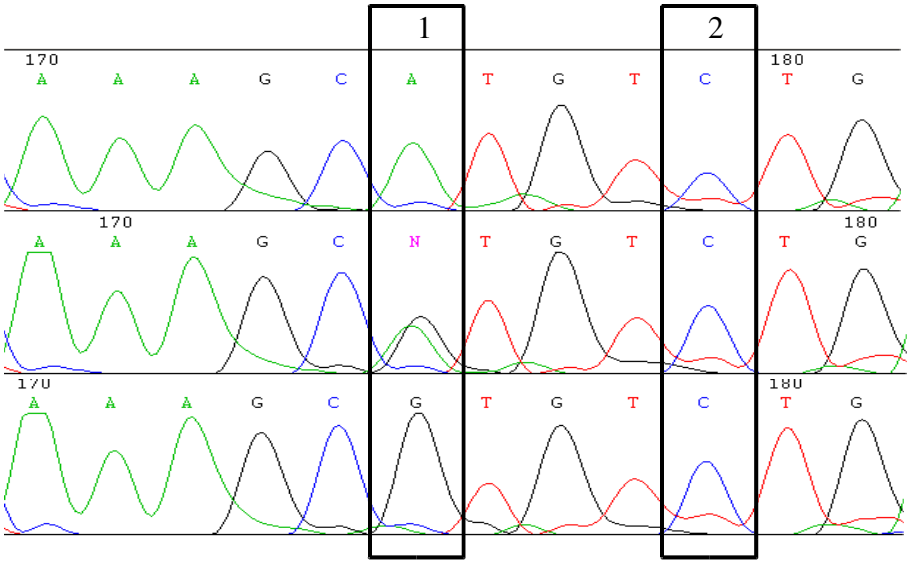


Figure 1. SNP discovery by alignment of sequence traces obtained from direct sequencing of genomic PCR products.
It is not always possible to distinguish between sequence artefacts and true polymorphism, when two peaks are present at one position. Box 1: top sequence homozygote AA, middle sequence heterozygote AG, bottom sequence homozygote GG. Box 2: The polymorphism detection software Polyphred [58] has considered the top and bottom sequences as heterozygote CT and the middle one as homozygote CC. Clonal sequence removes many of such ambiguities, since any double peak is a sequence artefact.

in which it is not always easy to distinguish between sequencing artefacts and polymorphism when double peaks, as expected in heterozygotes, are observed (Fig. 1). Therefore, different approaches based on the comparison of sequences obtained from cloned fragments can be considered for developing an SNP map of a genome. In this case, any double peak in a sequence trace is always considered as an artefact. The comparison of sequence data from EST production projects, especially if the libraries used were constructed using tissues from different individuals, can be a good source of SNPs that will have the additional interest of a greater chance of being in a coding region and hence have an influence on phenotypes [63]. Over a thousand SNPs have thus been identified in chickens [39]. However, the numbers generated by this approach will be limited, due to the selection pressure undergone by the coding sequences. In some rare instances, SNPs detected from EST sequence data, will in reality be a result of RNA editing. As a similar type of approach, in genomes for which complete genomic sequencing projects are undertaken, sequence comparisons in BAC clone overlaps will be a source of polymorphism.

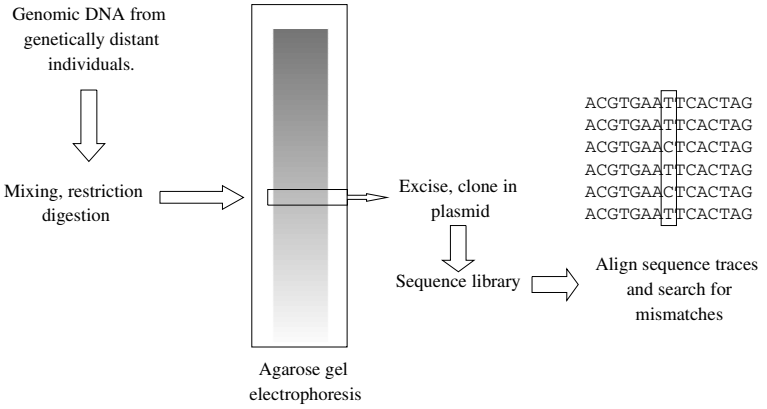


Figure 2. Reduced representation shotgun (RRS), for SNP discovery.
As a test for human SNP discovery, the *Bgl*II restriction enzyme was used. There is on average one *Bgl*II restriction site every 3 100 bp in the human genome, giving 26 000 fragments between 500 and 600 base pairs, representing 0.5% of the genome. Therefore, 52 000 sequences are needed for a twofold coverage. To develop high numbers of SNPs, The SNP Consortium (TSC) used several restriction enzymes and size ranges, to produce several libraries shared between sequencing centres [70].

The drawback in this case will be an uneven distribution of SNPs, due to the dependence of SNP detection on the number of overlapping BAC clones of different genetic origin along the genome [70]. These two approaches have the inconvenient of depending highly on the choice of the individuals at the origin of the cDNA or BAC libraries. More recently, a new approach, termed reduced representation shotgun (RRS) [3] was used for the production of a very high number of SNPs in humans. In this approach, DNA from different individuals are mixed together and plasmid libraries composed of a reduced representation of these genomes are produced by using a subset of restriction fragments purified by agarose gel electrophoresis (Fig. 2). A 2–5 fold shotgun sequencing of the libraries is performed and aligned overlapping sequences are screened for polymorphism. This last "*in silico*" step of identifying the SNPs in the sequence traces, whatever the way they were produced, has greatly benefited from the development of programs estimating the quality of base calling, such as PHRED [22, 23] and of other programs using this quality assessment for polymorphism detection, such as POLYPHRED [58] or POLYBAYES [56].

When searching for SNPs, care must be taken since there is the possibility of false positives due to the alignment of sequences from repeated loci, especially in random approaches such as RRS and the comparison of EST sequences. This can be partially overcome for species in which databases of repeated elements are available, that can be used to filter the sequence reads prior to alignment. However, the case of duplicated loci always remains difficult to manage.

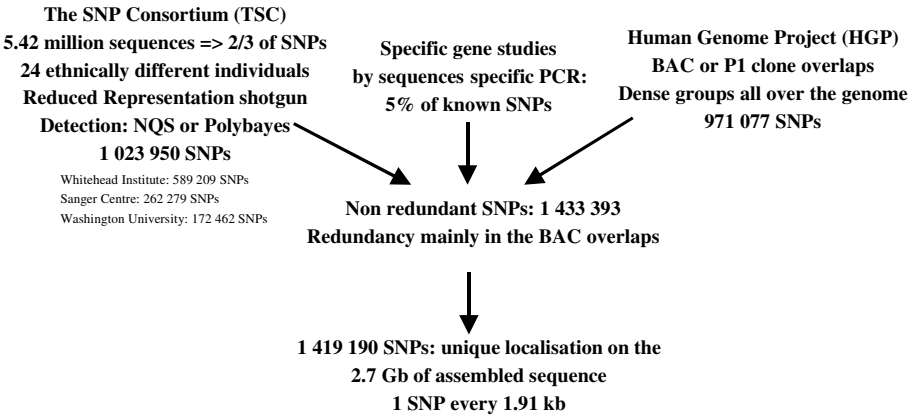


Figure 3. Generation of a 1 419 190 SNP map of the human genome. Over 2 million SNPs were generated by the reduced representation shotgun (RRS), by the analysis of clone overlaps from the Human Genome Project and by the analysis of specific genes. Localisation was performed by comparison to the assembled human genome sequence [70].

3.2. The human genome example

As often in molecular genetics, work progress in the human genome is the most advanced and an overview of what has been going on lately in this field will help understand what may be the future of animal genetics. Studies on numerous SNPs in defined regions, generally each concerning one gene, have been published with estimates of SNP frequencies and the extent of linkage disequilibrium. The involvement of specific SNP haplotypes in given phenotypes, usually diseases, has also been investigated. However, recently a more general approach in SNP development and analysis was followed.

High numbers of SNPs were generated by two main approaches. Shotgun sequencing of reduced representations of the genome, composed of a mixture of 24 ethnically diverse individuals [12], was performed by The SNP Consortium (TSC), composed of biotechnology and pharmaceutical companies (http://snp.cshl.org/). Also, a sequence comparison of regions of overlap between the large insert BAC (bacterial artificial chromosome) clones sequenced by the Human Genome Project (HGP) (Fig. 3) was done. By March 2001, 2.84 million SNPs had been deposited in a public database, 1.65 million of which were non-redundant [55]. Mapping of the SNPs was performed by sequence comparison with the assembled human genome sequence. In total, a map of 1.42 million SNPs, providing an average density of one SNP every 1.91 kb, was produced by February 2001 [70]. A few general conclusions can be withdrawn from this work, such as the normalised measure of heterozygosity (π), representing the likelihood that a nucleotide position will

be heterozygous, when compared across two chromosomes chosen at random from the population. For the human genome, $\pi = 7.51 \times 10^{-4}$, the expectation when comparing two chromosomes is therefore one SNP every 1 331 bp. With such high densities available, general detailed genome-wide studies can give new insights into population and genome dynamics. Although general studies on linkage disequilibrium (LD) show a heterogeneity between genomic regions, it extends on larger distances than first suspected in human populations, suggesting the occurrence of ancient demographic events, such as bottlenecks and migrations [65]. Genome dynamics can also be studied in great detail and for instance, the fine haplotype structure of human chromosome 21 was studied by determining the SNP content of 20 somatic cell hybrids, each containing a unique chromosome 21 of a different origin. More than 35 000 SNPs were thus identified, with known allelic phases and it was thus shown that large blocks of limited haplotype diversity exist on this chromosome [61]. Similar results indicating a structure composed of discrete blocks of 10 to 100 kb, each having only a limited number of common haplotypes and separated by small recombination hot spot regions, have been described in the class II region of the major histocompatibility complex [34] and over a 500 kb region of chromosome 5, in which 11 blocks of low haplotype diversity covered more than 75% of the sequence [17]. A study of 135 kb out of nine genes, has also revealed long stretches of linkage disequilibrium, suggesting that the common haplotype diversity of genes should be defined by a systematic approach, as an aid to the evaluation of their implication in common diseases [35]. However, if the long-range linkage disequilibrium induced by the underlying haplotype structure of the genome will help in defining small regions influencing traits in the first place, it will be difficult afterwards to pinpoint causal mutations on the basis of genetic evidence alone. Indeed, many SNPs will have equivalent association properties within a highly conserved common haplotype [66]. Association between a marker and a trait may even be difficult to find, in the case of a recent low frequency causal mutation embedded in a more ancient common haplotype.

3.3. Farm animals

No such extended studies have yet been made for farm animals, but from the limited data available, indications of high densities of SNPs in defined regions can be found. A sequencing study of fragments of the leptin and amyloid precursor protein (APP) genes in 22 diverse individuals from the two subspecies *Bos taurus* and *Bos indicus*, gave π values of 0.0026 and 0.019 respectively [41]. Within *Bos Taurus* alone, the π values were 0.0023 (one SNP every 434 bp) and 0.0096 (one SNP every 104 bp) for these fragments. Although it is clear from this study that the *APP* region studied is hypermutable, it can be concluded that high levels of diversity exist in this species. This has

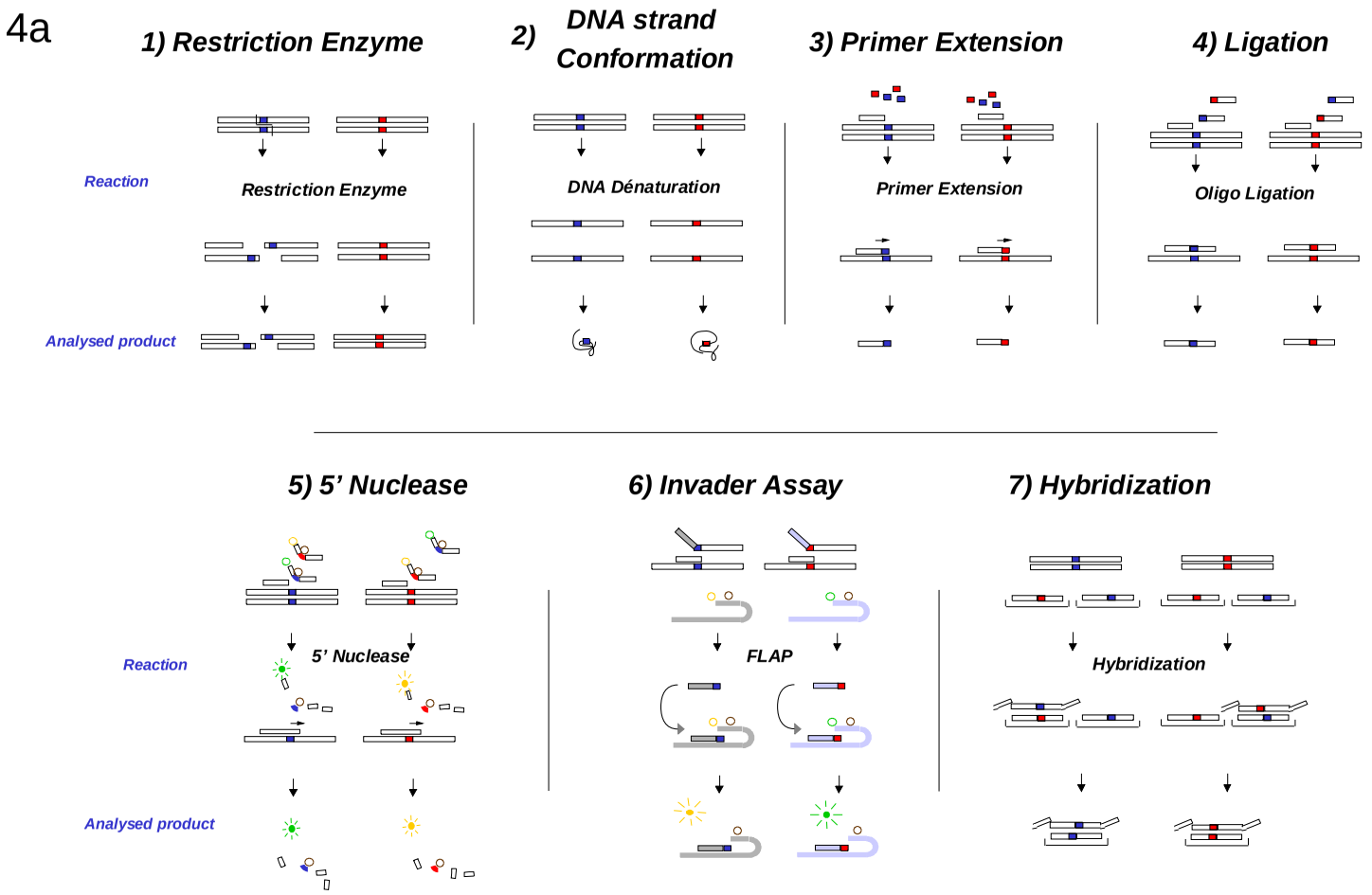
been confirmed by a study of 5.3 kb of genomic DNA from cytokine genes, in 26 individuals from a cattle reference population, in which an average 1 SNP per 443 bp was found [31]. These higher heterozygosity values found in cattle as compared to humans, may be a consequence of a pre-selection of the fragments studied, previously known to contain SNPs. However, studies in primates showing that diversity is reduced in humans, as compared to great apes [36], could suggest an alternative explanation for this phenomenon. In chickens, one SNP per 225 bp was observed in a survey of 31 000 bases analysed from broiler and layer lines [73] and one SNP per 2 119 bp was observed in chicken ESTs [39]. However, in these studies, the number of individuals sampled was not indicated and the heterozygosity value is therefore not available. A more random approach was also undertaken in chickens, in which a diversity study on more than 3 kb of DNA in 100 individuals from diverse European chicken breeds indicated varying levels of diversity ranging from no SNP to 17 SNPs in fragments of 500 bp each [79].

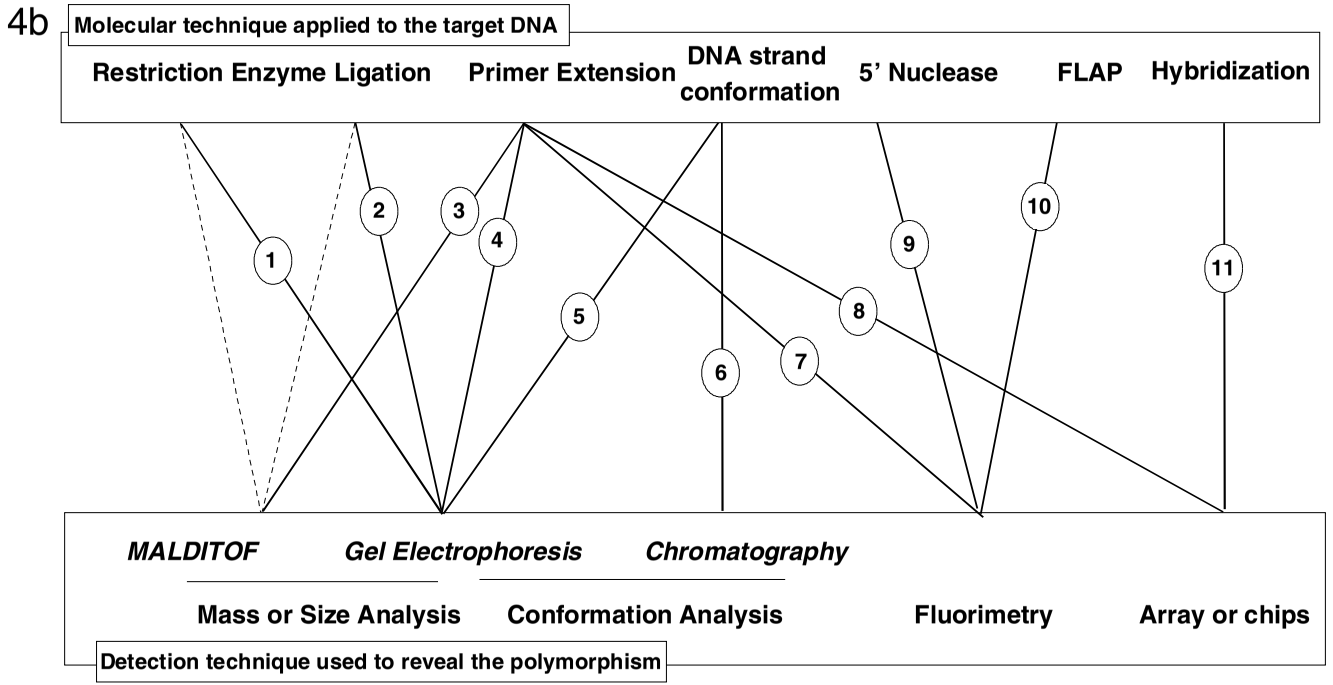
4. GENOTYPING SNPS

For microsatellite markers, there is a standard procedure for genotyping, involving PCR and size determination of the amplified fragment by acrylamide gel electrophoresis. The only differences in the techniques used in different laboratories are minor ones, principally concerning the use or not of an automatic sequencing machine for size determination. For SNP genotyping, this is not the case, and there are many techniques available. One key feature of most SNP genotyping techniques, apart from those based on direct hybridisation, is the two step separation: 1) generation of allele-specific molecular reaction products; 2) separation and detection of the allele specific products for their identification (Fig. 4). Due to the very broad range available, we will only present the main categories of SNP genotyping techniques here. Many are available as commercial kits.

4.1. Direct hybridisation techniques: from ASO to chips

Most hybridisation techniques are derived from the Dot Blot, in which DNA to be tested, either genomic, cDNA or a PCR reaction, is fixed on a membrane and hybridised with a probe, usually an oligonucleotide. In the Reverse Dot Blot technique, it is the oligonucleotide probes that are immobilised. When using allele specific oligonucleotides (ASOs), genotypes can be inferred from hybridisation signals. Throughput has now been greatly improved by using filters or glass slides containing very high probe densities. However, although conceptually simple, hybridisation techniques are error prone and need carefully designed probes and hybridisation protocols [59]. The latest


4a
1) Restriction Enzyme
2) DNA strand Conformation
3) Primer Extension
4) Ligation
Reaction
Restriction Enzyme
DNA Dénaturation
Primer Extension
Oligo Ligation
Analysed product
5) 5' Nuclease
6) Invader Assay
7) Hybridization
Reaction
5' Nuclease
FLAP
Hybridization
Analysed product



1 : *PCRRFLP* , **2** : *LAR or OLA*, **3** : *Good Assay,* **4** : *Minisequencing techniques, Snapshot ...,* **5** : *SSCP or DGGE*, **6** : DHPLC, **7** : Pyrosequencing, READit, **8** : *SNP it,* **9** : *Taqman*, **10** : *Invader Assay*, **11** : *Microarray or DNA chips*

Figure 4. SNP genotyping techniques.
4a: principal molecular reactions used to generate allele-specific signals.
4b: links between the signal generation and detection. The reason for the broad range of techniques available appears clearly, since many of the products generated by the allele-specific reactions can be detected with different methods.

improvements of this family of techniques, is the use of DNA chips, on which the probes are directly synthesised using a parallel procedure involving masks and photolithography [62]. The densities thus obtained are extremely high and reliability is improved by using a tiled array scheme, multiplying the number of probes used for each base position questioned [29,80].

4.2. Techniques involving the generation and separation of an allele-specific product

4.2.1. Restriction enzyme cutting

If the SNP to be studied involves a restriction enzyme site, PCR-RFLP can be a genotyping procedure that is easy to set up in any molecular biology laboratory. PCR products, if cut by the restriction enzyme, will generate typical fragments to be analysed by a size fractionation procedure, usually gel electrophoresis.

4.2.2. Single strand DNA conformation and heteroduplexes

Single strand conformation polymorphism (SSCP) is based on the specificity of folding conformation of single stranded DNA, when placed in non-denaturing conditions. One single base difference in DNA fragments of up to 300 bp, will usually change the conformation in a way that can be detected by non denaturing poly-acrylamide gel electrophoresis. Denaturing gradient gel electrophoresis (DGGE), is based on the fact that the melting point of double stranded DNA will be influenced by the presence of a mismatch. When the melting point is reached in a poly-acrylamide gel containing a gradient of denaturant, the electrophoretic mobility will be reduced. In a more recent version of this technique, denaturing high performance liquid chromatography (DHPLC), is used for the separation of the heteroduplex and homoduplex strands [51].

4.2.3. Primer extension

In this technique, an oligonucleotide is used, to prime DNA synthesis by a polymerase, as performed in a standard sequencing reaction or in PCR. Two main variations of the technique exist, the substrate being for both a PCR product obtained from the genomic DNA to be tested. In the first primer extension technique, two oligonucleotides are used, each with a $3'$ nucleotide complementary to one of the SNP alleles, since only perfectly matched oligonucleotides will prime DNA polymerase extension with dNTPs. One possibility for allele separation is to perform the primer extension directly on microarrays [60]. The use of mismatched primers can also theoretically be used to perform an allele-specific PCR, in which the oligonucleotides specific

for each allele are of different sizes or labelled with different dyes. However, in practice, the PCR conditions can be difficult to set up and reliability is low. In the second primer extension technique, a single base extension (SBE) primer is used, whose 3′ end is positioned on the base just preceding the SNP to be tested. The DNA polymerase is then used to incorporate labelled ddNTPs, each of four labelled with a different fluorescent dye. Any method that will allow to separate the labelled oligonucleotides from the non-incorporated ddNTPs, will be able to score the results simply on a fluorescence plate reader. Multiplexing of this procedure has been described thus reducing costs and improving throughput. In these methods, the different loci genotyped simultaneously are separated either by gel electrophoresis [49] or by hybridisation to arrayed tags [32]. Recently, Matrix-assisted laser desorption/ionisation time-of-flight mass spectrometry (MALDI-TOF) was developed as a tool for differentiating genotypes, by comparing the mass of DNA fragments after a single ddNTP primer extension reaction, in which no labelling is necessary. The precise mass of the product, that will depend on which ddNTP is incorporated, is determined. High levels of throughput and automation can be attained [8].

4.2.4. Oligonucleotide ligation assay (OLA)

Oligonucleotides are designed so that they join at the position of the polymorphism to be tested. Covalent joining, performed by a DNA ligase, occurs only when the match is perfect. The test is usually performed by designing two oligonucleotides specific for each allele and labelled differently on one side of the SNP, and one common oligonucleotide on the other. Detection of the alleles can be performed directly in the microplate wells by colorimetric approaches [76]. Multiplexing and the use of gel separation has also been described [28].

4.2.5. Pyrosequencing

Pyrosequencing is a recent rapid re-sequencing technology, in which template-mediated, oligonucleotide primed incorporation of nucleotides by a polymerase, is monitored by a measure of pyrophosphate (PPi) release. The four possible nucleotides are injected sequentially in the reaction mixture and the succession of successful incorporations, recorded on a pyrogram, gives the sequence. Comparison of the sequences with a reference enables to score SNPs [68]. An advantage of the method is that any new polymorphism will be detected. However, specific equipment is needed for the injection of the nucleotides.

4.2.6. Exonuclease detection (TaqMan)

The $5' \rightarrow 3'$ exonuclease activity of Taq polymerase is used to degrade an internal fluorescence resonance energy transfer (FRET) probe, that contains

a reporter and a quencher fluorescent dye. As long as they are linked to the oligonucleotide, the dyes are close together and the fluorescence is quenched. Upon degradation of the probe by the Taq polymerase, the fluorophore is released and the fluorescence thus emitted can be monitored. This reaction can be allele-specific, by using two different internal probes [46].

4.2.7. Invasive cleavage of oligonucleotide probes (invader assay)

This assay uses the property of Flap endonucleases (FENs), for removing the redundant portions (flap) from the 5′ end of a downstream DNA fragment overlapping an upstream (invader) DNA fragment. An invader oligonucleotide is designed, with its 3′ ending on the polymorphism to be tested. Two oligonucleotide signal probes are designed, overlapping the polymorphic site and each corresponding to one of the alleles. After displacement of the signal probes by the invader probe, FEN-mediated cleavage occurs only for the perfectly matched allele-specific signal probe [52]. Generation of the cleaved fragment is monitored, for instance by using it in a second reaction as an invader probe to cleave a fluorescence resonance energy transfer (FRET) probe [45]. This assay does not require PCR amplification of the locus to be tested and scoring is done using a simple fluorescence plate reader.

4.3. Changing genotyping techniques: the example of PrP

From the progress made in genotyping techniques, but also due to the number of different SNPs and/or individuals to genotype simultaneously, and the throughput needed, different options will be chosen. For example, the genotyping of polymorphisms at codons 136, 154 and 171 of the ovine PrP gene, implicated in susceptibility mechanisms to scrapies in sheep, was recently done at Labogena (Jouy-en-Josas, France) by using the PCR-RFLP method. After having improved throughput first by switching from an agarose gel method to a procedure using an automatic sequencing machine [4], the most recent genotyping set up for these 3 SNPs will now use the Taqman assay, based on 5′ → 3′ exonuclease removal of a quencher and fluorimetry (Boscher and Amigues, personal communication).

4.4. Which technique for the future?

It is difficult to predict if one technique, from the broad range available, will emerge in the future as a standard, especially since the needs will vary quite a lot between the extremes, such as the academic laboratory performing medium-scale studies, and commercial companies or genome centres aiming at very high throughput. In the first case, the choice may be influenced by the equipment and expertise available in the laboratory, whereas in the second,

to invest in new expensive dedicated machinery that can be necessary for some of the genotyping techniques, is less of a problem. Another important point to consider is the type of project envisaged, since it is quite different to perform genotypes with a limited number of SNPs on very large population samples, or a large number of SNPs on a limited number of individuals. In the first case, techniques that require an important investment in consumables specific to each SNP, such as expensive dedicated FRET primers, as used in the Taqman or Invader assays can be used, whereas they should be excluded in the second. Many techniques described here are protected by patents, a fact that can influence prices. Also, the scale of a study has an influence on the price of the consumables. Nevertheless, for studies involving large sets of samples, the use of primer extension techniques analysed by MALDI-TOF technology hold high promises in terms of automation, accuracy, throughput (a few seconds per genotype, for the acquisition step) and price (20 cents per genotype, Gut, personal communication). Pyrosequencing is also a very promising technique, with prices and throughput that might reach those of MALDI-TOF. It also has the great advantage of generating a complete short sequence stretch of about 50 base pairs, instead of just one genotype at a single base.

5. SNPS *VERSUS* MICROSATELLITES: ALLELE CALLING AND QUALITY OF DATA

5.1. Technical considerations

One technical problem with microsatellites is the fact that it is not always possible to compare data produced by different laboratories, due to the eventuality of inconsistencies in allele size calling. If this is usually not a problem for familial studies, such as those performed in QTL scans, it can be a serious issue when genotyping data from isolated individuals are used, such as in population studies. Such inconsistencies are mainly due to the large variety of automatic sequencing machines used, each providing different gel migration, fluorescent dyes and allele calling software possibilities. For instance, the type of fluorescent dye used will influence migration, and moreover, this will depend on the length and sequence of the DNA strand [30]. In some cases, even the use of multiple standard samples loaded on gels does not solve problems, particularly when large size differences between alleles exist.

Another particular case of error in size determination is due to the PCR reaction itself: according to the conditions used, the Taq polymerase catalyses the addition of an extra base (usually an adenine) at the $3'$ end of the PCR product. The proportion of fragments with this extra base may vary from none to 100%, inducing one base pair size differences and complicating data analysis. Although biochemical treatments after PCR or modification of PCR primers can circumvent this problem [9,26], they are seldom used.

Allele definition for microsatellites is done by assuming that size variation of PCR products is directly correlated with differences in repeat numbers of the simple motif. Although this is generally true, in some instances, size variations can be due to small deletions or insertions in flanking sequences and two PCR products of identical sizes can in reality be different alleles.

The allele nomenclature problem is much simpler in the case of SNPs, for which the results can just be coded as a YES/NO problem, in which each of the two alleles can be simply considered as being present or absent. This simplification in the scoring of alleles will enable the data analysis step of genotyping to be automated to a higher degree than with microsatellites, which still require a great investment of time for reading the data, even with the use of analysis software such as GENOTYPER (Applied Biosystems) or other automated allele analysis methods [33].

5.2. Statistical considerations

In any statistical analysis, one key point is the link between data and statistical treatment. The precise knowledge of the data generation process is needed in order to build a good statistical model. A particular point on which we would like to emphasize, is that of genotyping errors. Those inherent to human manipulation problems, can be overcome by careful planning of the laboratory procedures, the inclusion of well defined controls and increasing the degree of automation. However, those due to the biochemical processes used for genotyping are sometimes difficult to overcome and should be taken into account. The types of errors and the frequency at which they occur will be different between microsatellites and SNPs. In the case of microsatellites, the typical error will be that of size determination, in which case an allele will be replaced by one of the many other possibilities at the locus in consideration. In some instances, new alleles will be described, that are in reality artefacts. This can be easily corrected in family analyses, but the consequences of creating false alleles can be drastic in population genetics. In the case of SNPs, the only two frequent errors are the non detection of one of the two alleles, in which case a heterozygote individual will be genotyped as a homozygote, and the inverse, that is the false genotyping of a homozygote as a heterozygote. No creation of false alleles is possible. For both types of markers, the presence of null alleles is possible.

If the existence of typing errors is not taken into account, the results may be drastically biased and can be quite misleading. For instance, SanCristobal and Chevalet (1997) [71] showed in simulations of assignments of offspring to parents, that the assumption of the absence of typing errors can lead to a large number of wrong assignments even when only a few errors exist in reality in the data. Moreover, when a non null typing error rate is allowed for in the statistical treatment, even if higher than it really is, the assignment

process remains powerful. A demonstration of paternity assignments in red deer populations, taking genotyping errors into account, was done by Marshall *et al.* [54].

Likelihood-based approaches are generally powerful but not always robust. Statistical independence between markers is often required for simplicity of calculations. However, if too many markers are considered in an analysis, this assumption is obviously violated due to the limited size of genetic maps. The lower heterozygosity values of single locus SNPs as compared to microsatellites, implies the use of higher numbers and therefore raises the question of the statistical treatment of (at least partially) linked loci. If independence is nevertheless assumed, power is expected to fall down, and the estimates to be biased [43]. This is probably what happened to Ajmone-Marsan *et al.* (2001) [2] in a genetic diversity study in Italian goat populations, in which they reported that the coefficient of variation of the genetic indexes tested decreased only marginally when using more than 100 AFLP markers and bootstrapping on them. The use of alternative and model free methods, such as artificial neural networks (ANN) [6, 15], may circumvent this drawback in some cases, since they can give powerful results of assignment of individuals to a population, with the advantage that no hypotheses concerning the markers, and particularly the statistical independence, are needed. These methodologies should therefore enable the use of dense sets of SNPs.

Neutrality of markers is the base assumption in population genetics. The first idea that comes in mind is that microsatellites only seldom found in coding sequences, are by definition neutral, whereas in the case of SNPs, this will have to be checked for each marker. Indeed, even though most DNA sequences in eukaryotic genomes are non coding, many SNPs have been developed while working on specific genes or by comparison of EST sequences. However, the reality is not quite so clear cut and when tests for neutrality are performed, some microsatellites are clearly not neutral. Kantanen *et al.* (2000) [37], found that 2 out of 10 microsatellite loci significantly violated the null hypothesis of neutrality, when the Ewens-Watterson test was applied. In fact, this kind of approach can be used to test the effects of selection and was applied on a selection experiment in chickens, by calculating genetic distances between the initial and final generations, for many loci along a genetic map (Laval, personal communication). A marker presenting an increased genetic distance between generations, may suggest an effect of selection in its vicinity. Such methods could be used to detect regions containing QTLs in relation with the selection criteria.

Mutation can be neglected in population genetics problems involving small generation numbers, such as parentage testing and related problems, but also in genetic diversity studies of closely related breeds. Contrariwise, for high divergence times, mutation models are needed. Several possible models have

been proposed in the field of population genetics, and the choice of one or another has some influence on the statistical performances. For instance, Cornuet *et al.* [16], showed that genetic markers were always more efficient when evolving under the infinite allele model than under the stepwise mutation model, for selecting or excluding populations as the origin of individuals. Their inference was based on the genetic distance between individuals and populations. Authors seem to agree that microsatellite markers mutate according to a stepwise mutation model, whereas another model such as the infinite allele model will be used for SNPs.

Also, the much higher mutation rate of microsatellites, estimated to be as high as 1×10^{-5} [42] when compared to the 1×10^{-9} for SNPs [48,57], can be a concern, especially for association and linkage disequilibrium studies.

6. ON THE CHOICE OF MARKERS, ACCORDING TO SPECIFIC PROJECTS

Lets set aside the RFLP markers, mainly presented for their past importance. They are now replaced by PCR-RFLP so as to avoid using the Southern-blot technique; the various markers referred to in Tables I and II are still in use and the choice of one or another can be guided by the variety of parameters indicated, mainly according to the goal of the study and the importance of the species considered. Whatever the project, the higher heterozygosity values of microsatellites will enable to use lower numbers.

6.1. Traceability, paternity testing, population genetics

6.1.1. The use of fingerprinting techniques

In some species, only a limited number of microsatellite markers may have been produced, if any. In this case, the usual alternative is to use a fingerprinting technique, such as RAPD or AFLP. Although RAPD is technically less demanding than AFLP, the latter technique will produce more reproducible data, which will be easier to share between laboratories. The main interest of both techniques, is to use the same reagents, whatever the species studied. However, RAPD and AFLP produce bi-allelic dominant types of markers and therefore, to achieve the same resolution power as with microsatellites or even SNPs, a higher number of markers will have to be studied. Moreover, in most of the analyses performed, independence between markers is assumed and therefore, although fingerprinting techniques will easily produce high numbers of markers, care will have to be taken when using too many, especially since their map position is completely unknown.

Table I. The main categories of molecular markers.

Marker name	Variation type			Information content		
	SNP[1]	Indel[2]	VNTR[3]	2 dominant alleles	2 co-dominant alleles	Multi allelic co-dominant
RFLP	+	(+)[4]	(+)	−	+	(+)[5]
PCR-RFLP	+	(+)[4]	(+)	−	+	−
RAPD	+	(+)[4]	(+)	+	−	−
AFLP	+	(+)[4]	(+)	+	(+)[6]	−
SSCP	+	(+)[4]	(+)	−	+	(+)[5]
Microsatellite	−	(+)[7]	+	−	−	+
SNP	+	(+)[8]	−	−	+	−

[1] Single nucleotide polymorphism: any kind of base substitution. The fact that SNPs appear both as a variation type and a marker name, is due to the fact that in reality, many genotyping techniques used for genotyping SNPs are grouped under this generic marker name.
[2] Insertions and deletions.
[3] Variable number of tandem repeats.
[4] Although the RAPD, AFLP, RFLP, PCR-RFLP and SSCP techniques will detect base substitutions in the vast majority of cases, the two other types of DNA variation can also be analysed.
[5] In some instances, more than two alleles can be analysed.
[6] With an automatic sequencer, some markers can be scored as co-dominant.
[7] Variations in PCR product length can be due to a deletion in the sequence flanking the microsatellite.
[8] Many SNP detection techniques can also be used for scoring small insertions or deletions (indels).

6.1.2. SNPs versus *microsatellites*

Individual traceability of bovine meat

It has been proposed that standardised sets of SNPs could be used to produce digital DNA signatures for animal tagging [25]. After performing blind genotypings and allowing for a non-null error rate in the analyses, a minimal set of eight microsatellites could be kept, to assure perfect traceability of bovine meat [72]. Using this as a reference, a comparison with SNPs was done by drawing random bi-allelic markers assuming statistical independence, first with equal, then with uniformly distributed allelic frequencies. As expected, the presence of rare alleles leads to a dramatic fall in power, the maximum power being reached with (50%–50%) allelic frequencies. With uniformly distributed biallelic markers, a set of at least 30 was necessary to obtain perfect individual traceability (SanCristobal and Marimbordes, unpublished data).

Table II. Technical requirements and characteristics.

	Technical requirements				Technical characteristics			
Marker name	Restriction enzyme	PCR	Specific primers	Gel	Development effort	Genotyping effort	Reproducibility [1]	Accuracy [2]
RFLP	+	−	− [3]	+	High	High	High	Very high
PCR-RFLP	+	+	+	+	High	Medium	High	Very high
RAPD	−	+	−	+	Very low	Very low	Low	Very low
AFLP	+	+	−	+	Low	Very low	High	Medium
SSCP	−	+	+	+	Medium	Medium	Medium	Medium
Microsatellite	−	+	+	+	High	Low	High	High
SNP	−	+	+	+/− [4]	High	Variable [4]	High	Very high

[1] Refers to the genotyping error rate of the method: results may vary from one experiment to another.
[2] Refers to the precision at which true allele recognition can be performed.
[3] However, the RFLP technique relies on the use of a specific probe for the Southern-blot technique. Nowadays, RFLPs are usually genotyped by PCR-RFLPs, requiring specific primers.
[4] According to the genotyping technique used (see Fig. 3).

Parentage assignment, pedigree reconstruction and related problems

There are situations in animal breeding, in which relationships between two or more individuals, such as parent-child, full sibs, half-sibs, or unrelated individuals, have to be tested. Obviously, the size of the design will have an influence on the power of any statistical analysis performed. For instance, the assignment of the true father among a set of S sires is less powerful when S is large. Consequently, more loci will be needed to maintain a correct assignment level at a given rate when increasing S. Also, the number of loci will be critical if some individuals are missing, such as the mother or some of all of the potential fathers.

A numerical example can be given in the particular case where a finite set of potential sires is genotyped with markers and the mother is either or not genotyped. This situation is encountered in fish breeding, in which fry are mixed together so as to avoid environmental heterogeneity, or in sheep breeding, where several sires in natural mating systems are introduced together in a given flock. This latter case has been studied by considering 20 potential sires and using six microsatellite markers, with allelic frequencies estimated from French sheep breeds and assuming a 1% genotyping error rate. Simulation results indicate that 10% (respectively 18%) of assignment error occurs with the six microsatellites if the dam is typed (respectively non typed). Considering uniformly distributed allele frequencies, 30 (respectively 70) biallelic markers are needed to achieve the same error rates (SanCristobal and Amigues, unpublished data). This shows a higher need of biallelic markers (compared with multiallelic) when the mother is not genotyped, than when she is.

Even though the parent-offspring links are hence easy to ascertain with hypervariable loci, the grand-parent-offspring links require a greater genotyping effort: when matings are known, 95% correct grand-parentage assignments typically require at least twofold more alleles per locus than do 95% correct parentage assignments [47]. In terms of the number of loci, such a recognition may be prohibitive with SNPs.

When parentage is to be tested, typically one checks for incompatibilities between sire and child genotypes. Theoretically, no genotyping error is assumed to occur. The power of a set of markers depends on the global exclusion probability (*PE*). This kind of question is commercially important especially in horse breeding, but also in dairy cattle when sires have a high commercial value. Some formulas for *PE* can be found in Dodds *et al.* (1996) [19]. For instance, when a mother and child are genotyped, a wrong putative father is excluded with a probability $PE(k)$ at a locus with k alleles. Comparing a set of $L(2)$ loci with two alleles and a set of $L(k)$ loci with k alleles, the same exclusion probability is obtained if

$$[1 - PE(2)]^{L(2)} = [1 - PE(k)]^{L(k)}.$$

It follows that assuming equal allele frequencies at any locus, 2.23 (respectively 3.38) times more biallelic loci are needed than tri-allelic loci (respectively with four alleles).

Minimising the mean kinship between animals within populations has been suggested as a general approach for the conservation of genetic diversity. As for other applications, there is a strong effect of the polymorphism of markers on the good classification of the relatedness between individuals in categories such as full sibs, half sibs or parent-offspring pairs. For instance, almost twice as many loci of expected heterozygosity $He = 0.62$ (three alleles of frequencies 0.5, 0.3 and 0.2) are required to achieve the same accuracy as with loci of $He = 0.75$ (four alleles of frequency 0.25 each) [7]. To be able to accurately distinguish between non-inbred full sibs and half sibs, at least 30 to 50 unlinked markers with 5 to 10 alleles each are needed [21]. Therefore, the use of biallelic loci may not be a good solution for kinship estimation, since the numbers used will have to be very high.

The very high level of polymorphism found at some microsatellite loci in wild species, such as a 54 allele microsatellite locus found in the Pilot Whale by Amos *et al.* [5], is also in favour of their use: the development and especially the genotyping of SNPs representing a bigger effort, due to the larger numbers needed.

Genetic distances between populations

Genetic distances between pairs of populations are often the basis for diversity analyses. Usually, the simplest model is assumed, where a founder population splits into two daughter populations, which then diverge. For closely related populations, as encountered in diversity studies of livestock species, the meaningful parameter is the average inbreeding coefficient F. Distances estimating this inbreeding coefficient have approximately the following accuracy (see Laval 2001 for simulations and references) [44]:

$$\frac{2}{L(k_0 - 1)}\left(F + \frac{1}{m}\right)^2$$

where L is the number of loci, k_0 the number of founder alleles, and m the average number of sampled individuals in the final populations. It follows immediately that $(k_0 - 1)$ times more biallelic markers are needed to achieve the same genetic distance accuracy than a set of microsatellites with k_0 alleles.

This formula also implies that the coefficient of variation is very sensitive to the sample size for a small F, so the genotyping effort will have to be particularly important for very small divergence times, small sample sizes and when using bi-allelic markers, if accurate estimates of genetic distances are required.

When performing studies on admixture, bi-allelic loci provide little information about the admixture proportion and the time since admixture, even for very small amounts of drift, but they can be powerful when many loci are used [10].

Assignment of an individual to a population

The structure of populations, which varies according to the species studied due to variations in breeding strategies, will probably have an influence on the number of markers to be used for solving the problem. As few as four hypervariable microsatellite loci are sufficient to distinguish populations of brown trout and properly assign an individual to its population [6]. It has been shown in chickens, that this minimal number will have to be slightly larger, being between 10 and 20 [69]. However, given the low numbers of microsatellite markers needed, the matching number of SNP markers will remain low enough to develop a set with statistical independence.

6.2. Maps and QTL scans

Although the use of microsatellite markers is the best choice for the construction of a reference map for a species, the inclusion of type I markers (genes) is necessary both for the development of comparative maps and for the generation of positional candidate genes. Apart from a few cases in which microsatellites close to the coding sequences have been found, this has usually been done through the use of SNPs.

The mapping of regions containing QTLs involves the genotyping of markers covering the complete genome. These can be chosen from a reference map when available, in which case the markers are chosen as much as possible at regular intervals along the linkage groups. If available, the best to use in this case are the microsatellites, since they are highly informative and easy to use by PCR. However, in species for which no maps are available, QTL scans are performed with AFLP markers. Apart from the problem inherent to the use of dominant markers in this latter case, another drawback comes from the fact that no information on the position of the markers on the genome is available. Therefore, linkage groups specific to the cross studied have to be constructed, before the QTL analysis. After this step, since the polymorphism underlying an AFLP marker is usually an SNP, by converting the AFLP markers found linked to the QTL into a corresponding SNP [40], both advantages of co-dominance and locus-specificity will then be available. Also, once linkage to a QTL has been found, the real position on the genome will have to be determined by anchoring the particular linkage group on the cytogenetic map, so as to have means of developing new markers in a targeted way. Several approaches can then be taken, such as chromosome scraping or use of comparative mapping data.

One possibility for species of minor agricultural importance, for which mapping data is scarce, is to use as many locus-specific markers as possible, such as microsatellites or SNPs, together with the AFLPs. The former will help exchange mapping data between different crosses studied and also help provide information on the chromosomes concerned. The latter will help build linkage groups and ensure a correct genome coverage by allowing an increase in marker density.

It can be noted, that even in species for which dense maps with many microsatellites are available, the number of evenly spaced informative markers for a given cross may be too low. This will particularly be the case when closely related populations, such as two lines of a common origin, are divergently selected and crossed together for a QTL study. This is a current practice in chickens for instance. In such cases, SNPs are the only possible alternative.

6.3. Fine QTL mapping, candidate genes and complex traits

Several approaches can be taken for fine QTL mapping, such as increasing the number of meiosis events by increasing the size and/or the number of families for genotyping, selecting recombination events in recurrent backcrosses, using advanced intercross lines (AIL) or performing linkage disequilibrium and haplotype-based studies in outbread populations. However, whatever the approach taken, high densities of markers will be needed. In some instances, when the populations studied are closely related, even the microsatellite markers may not be heterozygous for the F1 animals. Also, for some species, such as chickens, the density of microsatellites will be low [64].

Testing of candidate genes and candidate polymorphisms in exons, promoters or other important regions such as splice sites, promoters or other regulatory regions, will have to be done using the SNP approach, since this will be the most common polymorphism and the more likely responsible for phenotypic variation.

When testing for the association between complex phenotypic traits and candidate loci, single-loci SNP analyses present a loss of information due to the bi-allelic nature of the markers, as compared to the multi-allelic microsatellites. However, by performing haplotype frequency estimations over several SNPs from a locus, this can be overcome [24] and even possibly improved, due to the fact that SNPs will more often be close to the site responsible for the variation than microsatellites.

7. CONCLUSION

Although in a strict molecular sense, SNPs are just what has been previously known as base substitutions, the fact of naming molecular markers by this

acronym meaning single nucleotide polymorphism, is an indication of the new importance that this type of polymorphism has in molecular genetics. Indeed, if in some instances, the lack of information due to the bi-allelic nature of SNPs is a limitation, there are cases in which they can provide valuable data on associations between specific genes or other DNA structures and phenotypes, or on population and genome dynamics.

The very high density of SNPs in genomes, usually allows to develop several of them in a single locus of a few hundred base pairs. By reconstructing haplotypes, multi-allelic systems can eventually be defined for analyses, to overcome the limitations due to the low heterozygosity of SNPs. With increasing progress being made in the molecular techniques used to produce SNP data, in the automation of allele scoring and in the development of algorithms for genetic analyses [1], the effort needed to produce an equivalent amount of information as with microsatellites may some day be equivalent.

REFERENCES

[1] Abecasis G.R., Cherny S.S., Cookson W.O., Cardon L.R., Merlin-rapid analysis of dense genetic maps using sparse gene flow trees, Nat. Genet. 30 (2002) 97–101.
[2] Ajmone-Marsan P., Negrini R., Crepaldi P., Milanesi E., Gorni C., Valentini A., Cicogna M., Assessing genetic diversity in Italian goat populations using AFLP markers, Anim. Genet. 32 (2001) 281–288.
[3] Altshuler D., Pollara V.J., Cowles C.R., Van Etten W.J., Baldwin J., Linton L., Lander E.S., An SNP map of the human genome generated by reduced representation shotgun sequencing, Nature 407 (2000) 513–516.
[4] Amigues Y., Mériaux J.-C., Boscher M.-Y., Utilisation de marqueurs génétiques en sélection : les activités de Labogéna, INRA Prod. Anim. Hors série (2000) 203–210.
[5] Amos B., Schlotterer C., Tautz D., Social structure of pilot whales revealed by analytical DNA profiling, Science 260 (1993) 670–672.
[6] Aurelle D., Lek S., Giraudel J.L., Berrebi P., Microsatellite and artificial neural networks: tools for the discrimination between natural and hatchery brown trout (*Salmo trutta* L.) in Atlantic populations, Ecol. Model. 120 (1999) 313–324.
[7] Blouin M.S., Parsons M., Lacaille V., Lotz S., Use of microsatellite loci to classify individuals by relatedness, Mol. Ecol. 5 (1996) 393–401.
[8] Bray M.S., Boerwinkle E., Doris P.A., High-throughput multiplex SNP genotyping with MALDI-TOF mass spectrometry: Practice, problems and promise, Hum. Mutat. 17 (2001) 296–304.
[9] Brownstein M.J., Carpten J.D., Smith J.R., Modulation of non-templated nucleotide addition by Taq DNA polymerase: primer modifications that facilitate genotyping, Biotechniques 20 (1996) 1004–1006, 1008–1010.
[10] Chikhi L., Bruford M.W., Beaumont M.A., Estimation of admixture proportions: a likelihood-based approach using Markov chain Monte Carlo, Genetics 158 (2001) 1347–1362.

[11] Collins D.W., Jukes T.H., Rates of transition and transversion in coding sequences since the human-rodent divergence, Genomics 20 (1994) 386–396.
[12] Collins F.S., Brooks L.D., Chakravarti A., A DNA polymorphism discovery resource for research on human genetic variation, Genome Res. 8 (1998) 1229–1231.
[13] Cooper D.N., Krawczak M., Cytosine methylation and the fate of CpG dinucleotides in vertebrate genomes, Hum. Genet. 83 (1989) 181–188.
[14] Cooperative Human Linkage Center (CHLC): Murray J.C., Buetow K.B., Weber J.L., Ludwigsen S., Scherpbier-Heddema T., Manion F., Quillen J., Sheffield V.C., Sunden S., Duyk G.M., Généthon: Weissenbach J., Gyapay G., Dib C., Morrissette J., Lathrop G.M., Vignal A., University of Utah: White R., Matsunami N., Gerken S., Melis R., Albertsen H., Plaetke R., Odelberg S., Yale University: Ward D., Centre d'étude du polymorphisme humain (CEPH) : Dausset J., Cohen D., Cann H., A comprehensive human linkage map with centimorgan density, Science 265 (1994) 2049–2054.
[15] Cornuet J.M., Aulagnier S., Lek S., Franck S., Solignac M., Classifying individuals among infra-specific taxa using microsatellite data and neural networks, C.R. Acad. Sci. III 319 (1996) 1167–1177.
[16] Cornuet J.M., Piry S., Luikart G., Estoup A., Solignac M., New methods employing multilocus genotypes to select or exclude populations as origins of individuals, Genetics 153 (1999) 1989–2000.
[17] Daly M.J., Rioux J.D., Schaffner S.F., Hudson T.J., Lander E.S., High-resolution haplotype structure in the human genome, Nat. Genet. 29 (2001) 229–232.
[18] Dib C., Fauré S., Fizames C., Samson D., Drouot N., Vignal A., Millasseau P., Marc S., Hazan J., Seboun E., Lathrop M., Gyapay G., Morissette J., Weissenbach J., A comprehensive genetic map of the human genome based on 5.264 microsatellites, Nature 380 (1996) 152–154.
[19] Dodds K.G., Tate M.L., McEwans J.C., Crawford A.M., Exclusion probabilities for pedigree testing farm animals, Theor. Appl. Genet. 92 (1996) 966–975.
[20] Donis-Keller H., Green P., Helms C., Cartinhour S., Weiffenbach B., Stephens K., Keith T.P., Bowden D.W., Smith D.R., Lander E.S., Botstein D., Akots G., Rediker K.S., Gravius T., Brown V., Rising M., Parker C., Powers J.A., Watt D.E., R. K.E., Bricker A., Phipps P., Muller-Kahle H., Fulton T.R., Ng S., Schumm J.W., Braman J.C., Knowlton R.G., Barker D.F., Crooks S.M., Lincoln S.E., Daly M.J., Abrahamson J., A genetic linkage map of the human genome, Cell 51 (1987) 319–337.
[21] Eding H., Meuwissen T.H.E., Marker-based estimates of between and within population kinships for the conservation of genetic diversity, J. Anim. Breed. Genet. 118 (2001) 141–159.
[22] Ewing B., Green P., Base-calling of automated sequencer traces using phred. II. Error probabilities, Genome Res. 8 (1998) 186–194.
[23] Ewing B., Hillier L., Wendl M.C., Green P., Base-calling of automated sequencer traces using phred. I. Accuracy assessment, Genome Res. 8 (1998) 175–185.
[24] Fallin D., Cohen A., Essioux L., Chumakov I., Blumenfeld M., Cohen D., Schork N.J., Genetic analysis of case/control data using estimated haplotype frequencies: application to APOE locus variation and Alzheimer's disease, Genome Res. 11 (2001) 143–151.

[25] Fries R., Durstewitz G., Digital DNA signatures: SNPs for animal tagging, Nat. Biotechnol. 19 (2001) 508.

[26] Ginot F., Bordelais I., Nguyen S., Gyapay G., Correction of some genotyping errors in automated fluorescent microsatellite analysis by enzymatic removal of one base overhangs, Nucleic Acids Res. 24 (1996) 540–541.

[27] Groenen M.A., Cheng H.H., Bumstead N., Benkel B.F., Briles W.E., Burke T., Burt D.W., Crittenden L.B., Dodgson J., Hillel J., Lamont S., de Leon A.P., Soller M., Takahashi H., Vignal A., A consensus linkage map of the chicken genome, Genome Res. 10 (2000) 137–147.

[28] Grossman P.D., Bloch W., Brinson E., Chang C.C., Eggerding F.A., Fung S., Iovannisci D.M., Woo S., Winn-Deen E.S., Iovannisci D.A., High-density multiplex detection of nucleic acid sequences: oligonucleotide ligation assay and sequence-coded separation, Nucleic Acids Res. 22 (1994) 4527–4534.

[29] Hacia J.G., Brody L.C., Chee M.S., Fodor S.P., Collins F.S., Detection of heterozygous mutations in BRCA1 using high density oligonucleotide arrays and two-colour fluorescence analysis, Nat. Genet. 14 (1996) 441–447.

[30] Hahn M., Wilhelm J., Pingoud A., Influence of fluorophor dye labels on the migration behavior of polymerase chain reaction–amplified short tandem repeats during denaturing capillary electrophoresis, Electrophoresis 22 (2001) 2691–2700.

[31] Heaton M.P., Grosse W.M., Kappes S.M., Keele J.W., Chitko-McKown C.G., Cundiff L.V., Braun A., Little D.P., Laegreid W.W., Estimation of DNA sequence diversity in bovine cytokine genes, Mamm. Genome 12 (2001) 32–37.

[32] Hirschhorn J.N., Sklar P., Lindblad-Toh K., Lim Y.M., Ruiz-Gutierrez M., Bolk S., Langhorst B., Schaffner S., Winchester E., Lander E.S., SBE-TAGS: an array-based method for efficient single-nucleotide polymorphism genotyping, Proc. Natl. Acad. Sci. USA 97 (2000) 12164–12169.

[33] Idury R.M., Cardon L.R., A simple method for automated allele binning in microsatellite markers, Genome Res. 7 (1997) 1104–1109.

[34] Jeffreys A.J., Kauppi L., Neumann R., Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex, Nat. Genet. 29 (2001) 217–222.

[35] Johnson G.C., Esposito L., Barratt B.J., Smith A.N., Heward J., Di Genova G., Ueda H., Cordell H.J., Eaves I.A., Dudbridge F., Twells R.C., Payne F., Hughes W., Nutland S., Stevens H., Carr P., Tuomilehto-Wolf E., Tuomilehto J., Gough S.C., Clayton D.G., Todd J.A., Haplotype tagging for the identification of common disease genes, Nat. Genet. 29 (2001) 233–237.

[36] Kaessmann H., Wiebe V., Weiss G., Paabo S., Great ape DNA sequences reveal a reduced diversity and an expansion in humans, Nat. Genet. 27 (2001) 155–156.

[37] Kantanen J., Olsaker I., Holm L.E., Lien S., Vilkki J., Brusgaard K., Eythorsdottir E., Danell B., Adalsteinsson S., Genetic diversity and population structure of 20 North European cattle breeds, J. Hered. 91 (2000) 446–457.

[38] Kappes S.M., Keele J.W., Stone R.T., McGraw R.A., Sonstegard T.S., Smith T.P., Lopez-Corrales N.L., Beattie C.W., A second-generation linkage map of the bovine genome, Genome Res. 7 (1997) 235–249.

[39] Kim H., Schmidt C.J., Decker K.S., Emara M.G. (2002). Chicken SNP discovery by EST data mining, in: Plant, Animal & Microbe Genomes X, 12–16 January 2002, San Diego. http://www.intl-pag.org/pag/10/abstracts/PAGX_P246.html

[40] Knorr C., Cheng H.H., Dodgson J.B., DNA cloning and sequence analysis of chicken AFLP, Anim. Genet. 32 (2001) 156–159.

[41] Konfortov B.A., Licence V.E., Miller J.R., Re-sequencing of DNA from a diverse panel of cattle reveals a high level of polymorphism in both intron and exon, Mamm. Genome 10 (1999) 1142–1145.

[42] Kruglyak S., Durrett R.T., Schug M.D., Aquadro C.F., Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations, Proc. Natl. Acad. Sci. USA 95 (1998) 10774–10778.

[43] Kuhner M.K., Beerli P., Yamato J., Felsenstein J., Usefulness of single nucleotide polymorphism data for estimating population parameters, Genetics 156 (2000) 439–447.

[44] Laval G., SanCristobal M., Chevalet C., Measuring genetic distances between breeds: use of some distances in various short term evolution models, Genet. Sel. Evol. accepted (2002).

[45] Ledford M., Friedman K.D., Hessner M.J., Moehlenkamp C., Williams T.M., Larson R.S., A multi-site study for detection of the factor V (Leiden) mutation from genomic DNA using a homogeneous invader microtiter plate fluorescence resonance energy transfer (FRET) assay, J. Mol. Diagn. 2 (2000) 97–104.

[46] Lee L.G., Connell C.R., Bloch W., Allelic discrimination by nick-translation PCR with fluorogenic probes, Nucleic Acids Res. 21 (1993) 3761–3766.

[47] Letcher B.H., King T.L., Parentage and grandparentage assignment with known and unknown matings: application to Connecticut River Atlantic salmon restoration, Can. J. Fish. Aquat. Sci. 58 (2001) 1812–1821.

[48] Li W.H., Gojobori T., Nei M., Pseudogenes as a paradigm of neutral evolution, Nature 292 (1981) 237–239.

[49] Lindblad-Toh K., Winchester E., Daly M.J., Wang D.G., Hirschhorn J.N., Laviolette J.P., Ardlie K., Reich D.E., Robinson E., Sklar P., Shah N., Thomas D., Fan J.B., Gingeras T., Warrington J., Patil N., Hudson T.J., Lander E.S., Large-scale discovery and genotyping of single-nucleotide polymorphisms in the mouse, Nat. Genet. 24 (2000) 381–386.

[50] Litt M., Luty J.A., A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene, Am. J. Hum. Genet. 44 (1989) 397–401.

[51] Liu W., Smith D.I., Rechtzigel K.J., Thibodeau S.N., James C.D., Denaturing high performance liquid chromatography (DHPLC) used in the detection of germline and somatic mutations, Nucleic Acids Res. 26 (1998) 1396–1400.

[52] Lyamichev V., Mast A.L., Hall J.G., Prudent J.R., Kaiser M.W., Takova T., Kwiatkowski R.W., Sander T.J., de Arruda M., Arco D.A., Neri B.P., Brow M.A., Polymorphism identification and quantitative detection of genomic DNA by invasive cleavage of oligonucleotide probes, Nat. Biotechnol. 17 (1999) 292–296.

[53] Maddox J.F., Davies K.P., Crawford A.M., Hulme D.J., Vaiman D., Cribiu E.P., Freking B.A., Beh K.J., Cockett N.E., Kang N., Riffkin C.D., Drinkwater R.,

Moore S.S., Dodds K.G., Lumsden J.M., van Stijn T.C., Phua S.H., Adelson D.L., Burkin H.R., Broom J.E., Buitkamp J., Cambridge L., Cushwa W.T., Gerard E., Galloway S.M., Harrison B., Hawken R.J., Hiendleder S., Henry H.M., Medrano J.F., Paterson K.A., Schibler L., Stone R.T., van Hest B., An enhanced linkage map of the sheep genome comprising more than 1000 loci, Genome Res. 11 (2001) 1275–1289.

[54] Marshall T.C., Slate J., Kruuk L.E., Pemberton J.M., Statistical confidence for likelihood-based paternity inference in natural populations, Mol. Ecol. 7 (1998) 639–655.

[55] Marth G., Yeh R., Minton M., Donaldson R., Li Q., Duan S., Davenport R., Miller R.D., Kwok P.Y., Single-nucleotide polymorphisms in the public domain: how useful are they?, Nat. Genet. 27 (2001) 371–372.

[56] Marth G.T., Korf I., Yandell M.D., Yeh R.T., Gu Z., Zakeri H., Stitziel N.O., Hillier L., Kwok P.Y., Gish W.R., A general approach to single-nucleotide polymorphism discovery, Nat. Genet. 23 (1999) 452–456.

[57] Martinez-Arias R., Calafell F., Mateu E., Comas D., Andres A., Bertranpetit J., Sequence variability of a human pseudogene, Genome Res. 11 (2001) 1071–1085.

[58] Nickerson D.A., Tobe V.O., Taylor S.L., PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing, Nucleic Acids Res. 25 (1997) 2745–2751.

[59] Pastinen T., Kurg A., Metspalu A., Peltonen L., Syvanen A.C., Minisequencing: a specific tool for DNA analysis and diagnostics on oligonucleotide arrays, Genome Res. 7 (1997) 606–614.

[60] Pastinen T., Raitio M., Lindroos K., Tainola P., Peltonen L., Syvanen A.C., A system for specific, high-throughput genotyping by allele-specific primer extension on microarrays, Genome Res. 10 (2000) 1031–1042.

[61] Patil N., Berno A.J., Hinds D.A., Barrett W.A., Doshi J.M., Hacker C.R., Kautzer C.R., Lee D.H., Marjoribanks C., McDonough D.P., Nguyen B.T., Norris M.C., Sheehan J.B., Shen N., Stern D., Stokowski R.P., Thomas D.J., Trulson M.O., Vyas K.R., Frazer K.A., Fodor S.P., Cox D.R., Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21, Science 294 (2001) 1719–1723.

[62] Pease A.C., Solas D., Sullivan E.J., Cronin M.T., Holmes C.P., Fodor S.P., Light-generated oligonucleotide arrays for rapid DNA sequence analysis, Proc. Natl. Acad. Sci. USA 91 (1994) 5022–5026.

[63] Picoult-Newberg L., Ideker T.E., Pohl M.G., Taylor S.L., Donaldson M.A., Nickerson D.A., Boyce-Jacino M., Mining SNPs from EST databases, Genome Res. 9 (1999) 167–174.

[64] Primmer C.R., Raudsepp T., Chowdhary B.P., Moller A.P., Ellegren H., Low frequency of microsatellites in the avian genome, Genome Res. 7 (1997) 471–482.

[65] Reich D.E., Cargill M., Bolk S., Ireland J., Sabeti P.C., Richter D.J., Lavery T., Kouyoumjian R., Farhadian S.F., Ward R., Lander E.S., Linkage disequilibrium in the human genome, Nature 411 (2001) 199–204.

[66] Rioux J.D., Daly M.J., Silverberg M.S., Lindblad K., Steinhart H., Cohen Z., Delmonte T., Kocher K., Miller K., Guschwan S., Kulbokas E.J., O'Leary S.,

Winchester E., Dewar K., Green T., Stone V., Chow C., Cohen A., Langelier D., Lapointe G., Gaudet D., Faith J., Branco N., Bull S.B., McLeod R.S., Griffiths A.M., Bitton A., Greenberg G.R., Lander E.S., Siminovitch K.A., Hudson T.J., Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to Crohn disease, Nat. Genet. 29 (2001) 223–228.

[67] Rohrer G.A., Alexander L.J., Hu Z., Smith T.P.L., Keele J.W., Beattie C.W., A comprehensive map of the porcine genome, Genome Res. 6 (1996) 371–391.

[68] Ronaghi M., Pyrosequencing sheds light on DNA sequencing, Genome Res. 11 (2001) 3–11.

[69] Rosenberg N.A., Burke T., Elo K., Feldman M.W., Freidlin P.J., Groenen M.A., Hillel J., Maki-Tanila A., Tixier-Boichard M., Vignal A., Wimmers K., Weigend S., Empirical evaluation of genetic clustering methods using multilocus genotypes from 20 chicken breeds, Genetics 159 (2001) 699–713.

[70] Sachidanandam R., Weissman D., Schmidt S.C., Kakol J.M., Stein L.D., Marth G., Sherry S., Mullikin J.C., Mortimore B.J., Willey D.L., Hunt S.E., Cole C.G., Coggill P.C., Rice C.M., Ning Z., Rogers J., Bentley D.R., Kwok P.Y., Mardis E.R., Yeh R.T., Schultz B., Cook L., Davenport R., Dante M., Fulton L., Hillier L., Waterston R.H., McPherson J.D., Gilman B., Schaffner S., Van Etten W.J., Reich D., Higgins J., Daly M.J., Blumenstiel B., Baldwin J., Stange-Thomann N., Zody M.C., Linton L., Lander E.S., Atshuler D., A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms, Nature 409 (2001) 928–933.

[71] SanCristobal M., Chevalet C., Error tolerant parent identification from a finite set of individuals, Genet. Res. Camb. 70 (1997) 53–62.

[72] SanCristobal M., Renand G., Amigues Y., Bosher M.-Y., Levéziel H., Bibé B., Traçabilité individuelle des viandes bovines à l'aide de marqueurs génétiques, INRA Prod. Anim. 13 (2000) 269–246.

[73] Schmid M., Nanda I., Guttenbach M., Steinlein C., Hoehn M., Schartl M., Haaf T., Weigend S., Fries R., Buerstedde J.M., Wimmers K., Burt D.W., Smith J., A'Hara S., Law A., Griffin D.K., Bumstead N., Kaufman J., Thomson P.A., Burke T., Groenen M.A., Crooijmans R.P., Vignal A., Fillon V., Morisson M., Pitel F., Tixier-Boichard M., Ladjali-Mohammedi K., Hillel J., Maki-Tanila A., Cheng H.H., Delany M.E., Burnside J., Mizuno S., First report on chicken genes and chromosomes 2000, Cytogenet. Cell Genet. 90 (2000) 169–218.

[74] Smith E.J., Shi L., Drummond P., Rodriguez L., Hamilton R., Ramlal S., Smith G., Pierce K., Foster J., Expressed sequence tags for the chicken genome from a normalized 10-day-old White Leghorn whole embryo cDNA library: 1. DNA sequence characterization and linkage analysis, J. Hered. 92 (2001) 1–8.

[75] Swinburne J., Gerstenberg C., Breen M., Aldridge V., Lockhart L., Marti E., Antczak D., Eggleston-Stott M., Bailey E., Mickelson J., Roed K., Lindgren G., von Haeringen W., Guerin G., Bjarnason J., Allen T., Binns M., First comprehensive low-density horse linkage map based on two 3-generation, full-sibling, cross-bred horse reference families, Genomics 66 (2000) 123–134.

[76] Tobe V.O., Taylor S.L., Nickerson D.A., Single-well genotyping of diallelic sequence variations by a two-color ELISA-based oligonucleotide ligation assay, Nucleic Acids Res. 24 (1996) 3728–3732.

[77] Vaiman D., Schibler L., Bourgeois F., Oustry A., Amigues Y., Cribiu E.P., A genetic linkage map of the male goat genome, Genetics 144 (1996) 279–305.
[78] Van Haeringen W.A., Den Bieman M., Gillissen G.F., Lankhorst A.E., Kuiper M.T., Van Zutphen L.F., Van Lith H.A., Mapping of a QTL for serum HDL cholesterol in the rabbit using AFLP technology, J. Hered. 92 (2001) 322–326.
[79] Vignal A., Monbrun C., Thomson P., Barre-Dirie A., Burke T., Groenen M., Hillel J., Maki-Tanila A., Tixier-Boichard M., Wimmers K., Weigend S. (2000) Estimation of SNP frequencies in European chicken populations, in: Conference Abstract Book of the 27th International Conference on Animal Genetics, 22–26 July 2000, Minneapolis, p. 71.
[80] Wang D.G., Fan J.B., Siao C.J., Berno A., Young P., Sapolsky R., Ghandour G., Perkins N., Winchester E., Spencer J., Kruglyak L., Stein L., Hsie L., Topaloglou T., Hubbell E., Robinson E., Mittmann M., Morris M.S., Shen N., Kilburn D., Rioux J., Nusbaum C., Rozen S., Hudson T.J., Lander E.S., Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome, Science 280 (1998) 1077–1082.
[81] Weber J.L., May P.E., Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction, Am. J. Hum. Genet. 44 (1989) 388–396.
[82] Weissenbach J., Gyapay G., Dib C., Vignal A., Morissette J., Millasseau P., Vaysseix G., Lathrop M., A second-generation linkage map of the human genome, Nature 359 (1992) 794–801.